Title
Evaluating the Utility of Multiple Trait Methods for Estimating Polygenic Risk Scores

Author
Fu, Jingyuan

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating the Utility of Multiple Trait Methods for Estimating Polygenic Risk Scores

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Jingyuan Fu

2020

ABSTRACT OF THE THESIS

Evaluating the Utility
of Multiple Trait Methods
for Estimating Polygenic Risk Scores

by

Jingyuan Fu
Master of Science in Bioinformatics
University of California, Los Angeles, 2020
Professor Eleazar Eskin, Chair

Polygenic risk score (PRS) is a method that utilizes the effect sizes of genetic variants on a particular disease or trait to evaluate an overall genetic risk for a certain individual. Such effect sizes are often estimated using traditional genome-wide association study (GWAS) for the trait of interest. There are methods developed that aim to improve the predictive power of PRS by incorporating the genetic information from multiple related traits. One existing popular method is MTAG, which requires GWAS summary statistics from multiple traits and is based on strong assumptions about genetic correlation across traits. We developed some variations of MTAG and evaluated their performance for computing PRS against GWAS, using a variety of trait data from UK Biobank as well as simulated data.

The thesis of Jingyuan Fu is approved.

Sriram Sankararaman

Jason Ernst

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2020

To my mother and father . . .

who have been patiently seeing my growth through all these years,

who are always supportive of every decision I made,

who nurtured and guided me, making me a good person.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my committee, who have been supporting my undergraduate and graduate study in this academic community. They have provided invaluable insights in my project and played an essential role in my decision to pursue my graduate study at UCLA. I would also like to extend my sincere appreciation to my mentor, Lisa Gai, who provided tremendous amount of knowledgeable support and valuable advice to this project, and who gave many perceptive comments on several drafts of my graduation thesis. I also wish to thank my partner in this project, Camille Huang, whose efforts and contribution are so essential that without them the completion of this thesis would not have been possible. I must also thank my friend and roommate, Huiling Huang, as well as all of my other friends, who provided so much help in my college years, both emotionally and academically, that pursuing my study here has been much smoother and happier that I previously imagined it would be. Lastly, I would love to thank all of my family members and my boyfriend, for all of their love and support.

CHAPTER 1

Introduction

Polygenic risk scores (PRS) have been commonly used for studying genetic architecture, and have gained increasing use for disease risk assessment [SHM16, LPE18]. They utilize the genetic information collected from studies such as GWAS, and take into account the estimated effects, whether small or large, of all of the single nucleotide polymorphism (SNPs) evaluated in the study.

The effect sizes used to compute the PRS are typically estimated using genome-wide association studies (GWAS), which fit a linear model for each SNP in the study on the trait. However, small effects are difficult to estimate accurately using traditional GWAS, sometimes requiring hundreds of thousands or millions of samples, in phenotypes that could be difficult to collect, such as phenotypes directly related to conditions of biological tissue samples.

People have been working on approaches to improve GWAS estimates by combining information from multiple related traits, as many pairs of traits exhibit genetic correlation, i.e. their effect sizes are correlated, and exhibit significant genetic correlation even in the absence of any significantly associated loci. Several existing methods leverage this genetic correlation in multiple traits to estimate variant effects from summary statistics [HLL17, MZL18, TW18, QC17]. In particular, the Multi-Trait Analysis of GWAS (MTAG) method has already been applied in a variety of settings [LTY17, GRA17].

However, MTAG may give biased effect size estimates when jointly analyzing large numbers of traits, due to limitations of the genetic correlation model, as it assumes the genetic

correlation across traits is identical across the genome, and that all SNPs have an effect in all traits. However, it has been shown that the genetic correlation between traits can vary from region to region [SMS17]. In cases where the assumption is violated, e.g. if a variant only has an effect in a subset of the traits, MTAG may overestimate the magnitude of the effect, with the degree of overestimation increasing with the number of traits in the analysis [TW18]. Hence, we present several variations of the MTAG method that model multiple-trait effects with more flexibility and could overcome the limitations of the MTAG method's assumption.

In this project, we assess the impact of estimating SNP effects from the MTAG model and its variations on PRS accuracy. We evaluate the multi-trait methods' utility for PRS in a variety of settings, using simulated data and real data from the UK Biobank (UKB). The phenotypes that we evaluate the models on include both anthropometric traits and psychiatric traits that have been previously used to test the MTAG model [TW18]. The end goal of this project is to have a comprehensive understanding of when multi-trait methods actually improve polygenic risk scores compared to traditional GWAS based approaches, with implications for improving phenotype prediction in a variety of settings.

CHAPTER 2

Methods

2.1 SNP filtering

We first introduce the SNP filter that we used in our analysis. We applied a SNP filter that was used by Turley *et. al* for SNP discovery and effect size estimates [TW18]. This filter was applied before running GWAS estimates. By applying the filter, we aim to filter out certain regions where the effect sizes are found to be strongly inflated, such as a neuroticism-associated inversion region in Chromosome 8.

2.2 Association testing and polygenic model for a single trait

We now describe how to perform a genome-wide association study (GWAS) at a SNP j on a single trait t, using data from N individuals. A traditional GWAS assumes the following linear model for the phenotype of individual i:

$$y_i = \beta_{jt} x_{ij} + e_i \tag{2.1}$$

where $e_i \sim N(0, \sigma_e^2)$.

Suppose we have a vector of standardized genotypes $\mathbf{x}$ at SNP j at each individual, and a vector of standardized phenotypes $\mathbf{y}$ for each individual. Then we may estimate the scalar effect β_j of SNP j on the phenotype t using linear regression:

$$\hat{\beta}_{jt} = \frac{1}{N}\mathbf{X}^\top \mathbf{y} \sim N\left(\beta_{jt}, \frac{1}{N}\sigma_e^2\right) = N\left(\beta_j, v\right) \tag{2.2}$$

where we have denoted $v^2 = \sigma_e^2/N$ for convenience. The variance σ_e^2 may be estimated as $\hat{\sigma}_e^2 = \frac{1}{N}(\mathbf{y} - \hat{\beta_{jt}}\mathbf{x}_j)^\top(\mathbf{y} - \hat{\beta_{jt}}\mathbf{x}_j)$.

GWAS assumes a linear model where only one SNP has non-zero effect, so the effects of any SNPs in LD with SNP j are also captured in β_j and subsequently in $\hat{\beta_{jt}}$. For this reason, β_{jt} is used here to refer to the marginal effect of SNP j. The software that we used for running GWAS estimates is PLINK (version 1.9) [PNT07], [Pur20].

Next, we describe the additive polygenic model used to compute a polygenic risk score (PRS). In this model, an individual's phenotype is simply the weighted sum of their standardized genotype at a set of SNPs, say a set of M independent SNPs. Recall that each β_j has been estimated on standardized phenotypes. Then the (standardized) phenotype for individual i is given by

$$y_i = \sum_{j=1}^{M} \beta_{jt} x_{ij} + e_i \tag{2.3}$$

where $e_i \sim N(0, \sigma_e^2)$ is Gaussian noise. It is also assumed that $\sum_{j=1}^{M} \beta_j \sim N(0, \sigma_g^2)$, with narrow heritability $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$ corresponding to the fraction of phenotypic variance explained by additive SNP affects.

To estimate the PRS for an individual i from a set of M independent SNPs, we compute

$$\hat{y}_i = \sum_{j=1}^{M} \hat{\beta}_{jt} x_{ij} \tag{2.4}$$

There are several options for choosing the set of SNPs to use in the PRS. For example, one may use all genotyped SNPs, SNPs that are GWAS-significant, or some LD-pruned subset of SNPs. If using non-independent SNPs as predictors for the PRS, we must apply a correction for LD so that the summation is over the non-marginal SNP effects. Otherwise, SNP effects will effectively be counted multiple times in the summation. There are existing softwares that perform this correction when computing the PRS.

2.3 Multivariate normal model for effects across multiple traits

Suppose that we have GWAS summary statistics from K traits at SNP j. This consists of GWAS estimated effect sizes $\hat{\beta}_{\mathbf{j}} = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK})$, as well as the sample variance of these estimates.

Also suppose that the true SNP effects across traits $\beta_{\mathbf{j}} = (\beta_{j1}, \ldots, \beta_{jK})^\top$ are drawn from a multivariate normal (MVN), such that

$$\beta_{\mathbf{j}} \sim N(\mathbf{0}, \mathbf{\Omega}) \tag{2.5}$$

where $\mathbf{0}$ is a vector of all zeros of appropriate dimension and $\mathbf{\Omega}$ is the genetic covariance matrix, such that entry ω_{ij} is proportional to the genetic correlation between traits i and j. Note that $\beta_{\mathbf{j}}$ is a vector of marginal SNP effects, that is, it includes the effects of other SNPs in LD with the SNP of interest.

Suppose we then have K studies, one for each trait. Given the true effects, the GWAS linear estimator will come from the following distribution

$$\hat{\beta}_{\mathbf{j}} | \beta_{\mathbf{j}} \sim N(\beta_{\mathbf{j}}, \mathbf{\Sigma}_j) \tag{2.6}$$

where $\mathbf{\Sigma}_j$ is the variance-covariance matrix for the estimated effects across studies for SNP j, or to put in other words, the variance-covariance matrix for estimation errors for SNP j. For each pair of traits (t, s), the entry $\Sigma_{t,s}$ is proportional to the environmental correlation across studies for this pair of trait, which may be non-zero if there is sample overlap across studies. In practice, we estimate $\hat{\mathbf{\Sigma}}_j$ using the following steps done by Turley *et. al* [TW18]. For a trait t with a study size of N_t at SNP j, $\hat{\Sigma}_{t,t,j} = \sigma^2_{e_t}/N_{t,j}$. For a set of two traits t, s with study sizes of N_t and N_s respectively at SNP j, $\hat{\Sigma}_{t,s,j} = \sigma_{e_t}\sigma_{e_s}/\sqrt{N_{t,j}N_{s,j}}$. Note that the study size N's could vary across SNPs for any specific trait.

Here, $\sigma^2_{e_t}$ and $\sigma_{e_t}\sigma_{e_s}$ are variance and covariance due to a sample overlap, that can be estimated from LD score regression intercepts [BF15].

In addition to LD score regression covariance intercepts that are used in the multi-trait analysis, we also collected genetic covariance, genetic correlation and heritability of each trait using LD score regression.

In practice, the variance-covariance matrix of effect size distribution, $\hat{\mathbf{\Omega}}$, is estimated using methods of moments described in Turley et. al [TW18].

$$\hat{\mathbf{\Omega}} = \frac{1}{M}\sum_{j=1}^{M}\left(\hat{\beta}_{\mathbf{j}}\hat{\beta}_{\mathbf{j}}' - \hat{\mathbf{\Sigma}}_j\right) \tag{2.7}$$

where $\hat{\beta}_{\mathbf{j}}$ is the GWAS estimates of effect sizes for SNP j.

2.4 Multi-trait estimator for effect size, assuming identical genetic covariance across the genome

Here we present the multi-trait linear estimator derived by Turley *et al.* for estimating effects from GWAS summary statistics, assuming all of the effect sizes share the same variance-covariance matrix $\mathbf{\Omega}$ [TW18].

Suppose we have GWAS summary statistics for a trait of interest and additional related traits, as well as covariance matrices $\mathbf{\Omega}$ and $\mathbf{\Sigma}$. For purposes of the derivation, we assume that the true values of $\mathbf{\Omega}$ and $\mathbf{\Sigma}$ are known. We call the trait of interest the primary trait, and the additional traits as auxiliary traits. Denote the GWAS effect sizes from K traits as $\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jK}$. [TW18] derive an estimator b_{MTAG} for the effect in the primary trait β_1:

$$b_{MTAG} = \frac{\left(\frac{\omega_1}{\omega_{11}}\right)^{\top}\left(\mathbf{\Omega} - \frac{1}{\omega_{11}}\omega_1\omega_1^{\top} + \mathbf{\Sigma}_j\right)^{-1}\hat{\beta}}{\left(\frac{\omega_1}{\omega_{11}}\right)^{\top}\left(\mathbf{\Omega} - \frac{1}{\omega_{11}}\omega_1\omega_1^{\top} + \mathbf{\Sigma}_j\right)^{-1}\left(\frac{\omega_1}{\omega_{11}}\right)} \tag{2.8}$$

Note that MTAG estimates the effects of all traits in the analysis jointly, but for consistency with other methods we test, we use the primary/auxiliary trait notation.

2.5 Two component mixture models for effects across traits

In the previous sections, we assume the true SNP effects are drawn from the same distribution across the genome, so that all SNPs have some effect in all traits. We now consider models where the true SNP effects across traits are drawn from a Gaussian mixture model with two components.

Let $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})^\top$ denote the effects of SNP j in traits $1, \ldots, K$. Let π_0 and π_1 be the mixing weights of each component, such that $\pi_0 = 1 - \pi_1$, and let γ_j be a latent variable for which component SNP j was drawn from. Say the trait of interest t_1 is the first entry in β, which we will refer to the primary trait, and other traits as auxiliary traits. The generative model for the true effects at a SNP is as follows.

$$\begin{aligned}
\gamma_j &\sim \text{Bernoulli}(\pi) \\
\beta_j|\gamma_j = 0 &\sim N(\mathbf{0}, \mathbf{\Omega}_0) \\
\beta_j|\gamma_j = 1 &\sim N(\mathbf{0}, \mathbf{\Omega}_1)
\end{aligned}$$

where $\mathbf{\Omega}_0$ is a K by K matrix chosen from one of the options described below, and $\mathbf{\Omega}_1$ is the genetic covariance matrix,

$$\mathbf{\Omega}_1 = \begin{bmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \cdots & \rho_{1k}\tau_1\tau_k \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\tau_1\tau_k & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

$\mathbf{\Omega}_1$ corresponds to the classic polygenic model with full genetic correlation, estimated using Eq. 2.7, while $\mathbf{\Omega}_0$ contains a subset of entries of $\mathbf{\Omega}_1$ reflecting our assumptions about the relationships between traits in the analysis. The mixing parameters π are estimated by expectation-maximization (EM).

We test four options for $\mathbf{\Omega}_0$ each corresponding to a different possible GMM, described below.

- If we assume the primary trait has an effect and there is no effect among other traits (we will refer to the corresponding model as GMM_A):

$$\boldsymbol{\Omega}_{0a} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \epsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon \end{bmatrix}$$

- If we assume the primary trait is sometimes sparse, but correlation amongst the auxillary traits is the same for all SNPs (GMM_B):

$$\boldsymbol{\Omega}_{0b} = \begin{bmatrix} \epsilon & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

- If we assume there is no sparsity in effect sizes, but the primary trait is sometimes uncorrelated with the auxillary traits (GMM_C):

$$\boldsymbol{\Omega}_{0c} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & \rho_{2k}\tau_2\tau_k \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \rho_{2k}\tau_2\tau_k & \cdots & \tau_k^2 \end{bmatrix}$$

- If we assume there is no sparsity in effect sizes, but effect sizes are sometimes completely independent in all traits (GMM_D):

$$\boldsymbol{\Omega}_{0d} = \begin{bmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_k^2 \end{bmatrix}$$

2.5.1 Model fitting

The mixture weights for these GMMs are estimated using Expectation-Maximization (EM). We describe the procedure here. Assuming we are given $\hat{\beta}_j$ for $j = 1, \ldots, M$. We use $N(\hat{\beta}_j|\mathbf{0}, \mathbf{A})$ to denote the density at $\hat{\beta}_j$ of a centered multivariate normal distribution with covariance matrix $\mathbf{A}$.

In the E step, we compute membership assignment for each SNP based on our current estimates for the mixture weights.

$$p(\gamma_j = 0|\hat{\beta}_\mathbf{j}, \mathbf{\Omega}_0, \mathbf{\Sigma}_j) \leftarrow \frac{N(\hat{\beta}_\mathbf{j}|\mathbf{0}, \mathbf{\Omega}_0 + \mathbf{\Sigma}_j)}{\sum_{c=0}^{1} N(\hat{\beta}_\mathbf{j}|\mathbf{0}, \mathbf{\Omega}_c + \mathbf{\Sigma}_j)\pi_c} \tag{2.9}$$

for $j = 1, \ldots, M$.

In the M step, we total the fraction of SNPs assigned to each component to get the mixture weights. The M step becomes

$$\pi_0 \leftarrow \frac{\sum_{j=1}^{M} p(\gamma_j = 0|\hat{\beta}_\mathbf{j}, \mathbf{\Omega}_0, \mathbf{\Sigma}_j)}{M} \tag{2.10}$$

for $j = 1, \ldots, M$.

We alternate between the E and M steps until convergence. These EM updates assume that each SNP is an independent sample from the mixture distribution, so they should be computed over an LD-pruned subset of all available SNPs.

2.6 Stratifying SNPs by LD score and MAF

The four GMMs in the previous section base both covariance components on the GWAS estimated effect sizes and covariance intercepts across all available SNPs. As an alternative approach, we also tested a model where each component corresponds to a different subset of SNPs, stratified by minor allele frequency (MAF) and LD score. Rare SNPs tend to have a

different effect size distribution than common SNPs, and SNP marginal effects include the effects of any SNPs in strong LD.

We tested a model with 4 bins. In this model, we split SNPs into top 50% or bottom 50% by LD score, and MAF < 0.1 or MAF ≥ 0.1, and split the summary statistics according to the bins. For each bin, we munged the corresponding summary statistics and obtained estimates of genetic covariance, covariance intercepts, genetic correlation and heritability using LDSC separately [BL15, BF15]. Then, we applied MTAG to each bin. We estimated the effect size variance-covariance matrix separately for each bin using Eq. 2.7, except now we are only using SNPs in that bin rather than using all of the SNPs in the analysis. We then computed the MTAG effect size estimators using Eq. 2.8 for each bin, and combined those effect size estimators for all of the SNPs.

2.7 Computing polygenic risk scores with correction for LD

In Eq. 2.3, we assumed we had an independent set of SNPs to compute the PRS. However in practice, there is widespread correlation between SNPs in the genome, i.e. linkage disequilibrium (LD). To account for LD, one may either select a subset of approximately independent SNPs using LD pruning and thresholding by GWAS p-value, or adjust the marginal SNP effects for LD before computing PRS.

We use LDpred (version 1.06) to adjust estimates of marginal SNP effects for LD before computing the PRS [VYF15], as in [TW18]. We use a random sample of 5,000 individuals as an LD reference panel. We set the LD radius to be 150 when generating LDpred SNP weights. We assume an infinitesimal model in LDpred, meaning that the fraction of causal SNPs assumed by LDpred is 1. We apply LDpred to GWAS, MTAG, and our five additional models for the anthropometric and psychiatric traits.

2.8 Validating method performance using simulated effect sizes from bivariate normal distribution

To validate the multi-trait methods' performance on simulated datasets, we adopted similar simulation settings to Turley et. al [TW18], except that we only generated effect sizes from a bivariate normal distribution model rather than a spike-and-slab model. In short, we generated 100,000 length-two "true" effect size vectors $\beta_{\mathbf{j}}$ using the bivariate normal distribution setting. The distribution has variance one and correlation r_β, and the effect sizes are drawn from this distribution with mean zero and variance covariance matrix $\mathbf{\Omega}_{\text{normal}}$ using 2.11. We also generated z-statistics, $\mathbf{Z}_j$, where for each SNP j, we add estimation errors to the true effect size vectors, and then divide the estimates by the standard deviation of estimation error. The estimation error is drawn from a bivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{\Sigma}$, which is calculated using Eq. 2.12. We then generated an estimate of $\mathbf{\Sigma}$, $\hat{\mathbf{\Sigma}}$, by adding independent, normally distributed noise to $\mathbf{\Sigma}$. We then estimated $\hat{\mathbf{\Omega}}$ using the methods of moments procedure in Eq. 2.7 and generated MTAG estimates and standard errors for each SNP and for each trait. We computed GMM estimates using GMM_A method.

$$\mathbf{\Omega}_{\text{normal}} = \begin{bmatrix} 1 & r_\beta \\ r_\beta & 1 \end{bmatrix} \tag{2.11}$$

$$\mathbf{\Sigma} = \mathbf{C}\mathbf{\Sigma}_{LD}\mathbf{C} \tag{2.12}$$

where $\boldsymbol{C} = \begin{bmatrix} \sqrt{\frac{1}{\chi_1^2 - 1}} & 0 \\ 0 & \sqrt{\frac{1}{\chi_2^2 - 1}} \end{bmatrix}$ and $\mathbf{\Sigma}_{LD} \equiv \begin{bmatrix} 1 & r_\varepsilon \\ r_\varepsilon & 1 \end{bmatrix}$

Here, we fixed r_ε to be 0.3, where r_β, χ_1^2 and χ_2^2 were changed in different settings. This equation is to capture the effect of estimation error on $\hat{\mathbf{\Sigma}}$ [TW18].

CHAPTER 3

Results

3.1 UK Biobank Results

3.1.1 Genetic correlation and trait selection

3.1.1.1 UK Biobank anthropometric traits

We chose a number of anthropometric traits in the UK Biobank and estimated the genetic correlation matrix using LD score regression. The genetic correlation matrix is estimated using in total 225268 SNPs, shown in Table. 3.1.

We selected the four sets of traits that we proceeded to analyze. The sets were (1) arm fat percentage in left arm, trunk fat percentage and waist circumference; (2) automated pulse measurement and pulse rate (during blood-pressure measurement); (3) automated pulse measurement and standing height; and (4) standing height, seated height, and systolic blood pressure. Note that we refer to the "pulse rate (during blood-pressure measurement)" trait as manual pulse for convenience. These four sets of traits are chosen to represent four different scenarios in multi-trait analysis respectively: strong positive genetic correlation for three traits with large datasets (set 1), strong positive genetic correlation for two traits with varying sizes of datasets (set 2), weak genetic correlation for two traits with large datasets (set 3), and a combination of strong and weak correlations with varying sizes of datasets (set 4) (Table. 3.1) [BF15]. The dataset sample sizes are shown in later sections (Table. 3.3).

3.1.1.2 UK Biobank psychiatric traits

Besides analyzing anthropometric traits in UK Biobank (UKB), we also analyzed the same set of traits as Turley *et. al* have analyzed in their paper [TW18]. They analyzed three psychiatric traits: depression, neuroticism and subjective well-being, using several large cohorts of data from UK Biobank, 23andMe and SSGAC. As we are only using UK Biobank traits, our samples do not have a large proportion of overlap with their samples. Therefore, we compared GWAS estimated effect sizes that we computed using UK Biobank dataset and the GWAS effect size estimates generated from the dataset that Turley group used. After testing for correlation (Fig. 5.1A and 5.1B), we decided to use only depression and neuroticism for our evaluation, as the GWAS estimated effect sizes of subjective well-being dataset did not correlate well with those computed by Turley group. It could be because there are multiple measures of subjective well-being, and 23andMe dataset uses different measures than UKB, making the UKB dataset alone not representative of the whole cohort in their analysis.

We continued to compute the genetic correlation between depression and neuroticism, which is 0.8079 (with a standard error of 0.0251), estimated using in total 225268 SNPs.

In future sections, we use abbreviations of the trait names to represent the actual trait for both anthropometric and psychiatric traits. The corresponding traits and descriptions are in (Table. 3.2)

3.1.2 Sampling data and stratifying SNPs

All traits used in these experiments had on the order of 3 to 4 million individuals with non-missing phenotype values, except for manual pulse and systolic blood pressure, which had on the order of 50k individuals. For the following experiments, we only used white British individuals in the UK Biobank. We randomly subsampled 200k individuals from the UK Biobank to use as a GWAS cohort, and 10k individuals for whom we computed PRS from

the different methods, such that there was no overlap between the two cohorts. Note that we did not take missing phenotypes into account when sampling, so the final GWAS sizes are proportionate the the original sample sizes. We used LDpred to estimate PRS from effect sizes [VYF15],[TW18].

After excluding any missing data from the samples, the number of final individual counts for each trait is in Table. 3.4 and 3.3, where systolic blood pressure and manual pulse have ¡20k samples, while the other traits have 190k to 200k samples.

After we stratify the SNPs into 4 bins using LD score and MAF, the number of SNPs in each bins are shown in Table. 3.5, and the gene correlation matrices and heritability estimates are shown in supplement (Table. 5.1, 5.2, 5.3, 5.4, 5.5)

3.1.3 Evaluation of PRS predictions

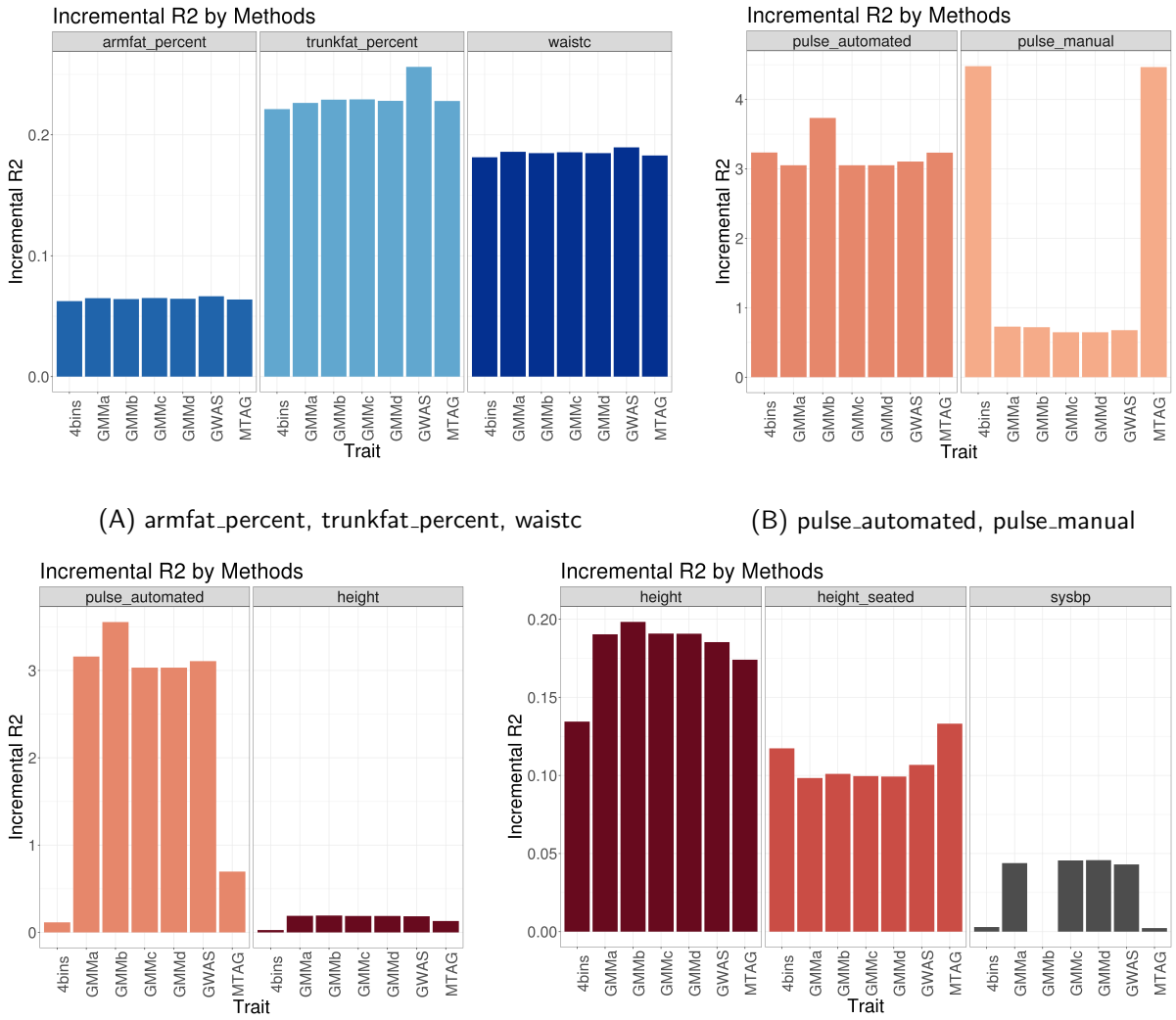
We applied GWAS, MTAG, four GMM models, and the 4 bin method to four sets of anthropometric traits and one set of psychiatric traits in the UK Biobank chosen to represent different scenarios for multi-trait analysis.

We then computed the Pearson correlation between the PRS from each method and the true phenotypes of the PRS cohort. We computed incremental r^2, the proportion increase in r^2 between the PRS estimated using the seven models, compared to prediction using linear model with covariates only (Fig. 3.1). We also computed incremental adjusted r^2, incremental r^2 corrected for additional number of predictors compared to covariate-only linear model (Fig. 3.2). We found that GMMs A-D typically performed similarly to each other, with GMM_B sometimes outperforming other GMM methods. The 4 bin model's performance also varied dramatically in different traits. We also noted that MTAG was consistently not the worst, and that its performance relative to GWAS was strongest for predicting manual pulse using automated pulse (Fig. 3.2B), though it did underperform GWAS for both traits in the pulse and height set. We suspect that MTAG is mainly useful

in situations where GWAS in the trait of interest is underpowered, but otherwise performs comparably to GWAS, or worse if the traits are weakly correlated.

We also computed the same metrics for the set of psychiatric traits that we are evaluating our methods on (Fig. 3.3).

The genetic correlation matrices calculated in the 4 bins model are listed in Supplement (Table. 5.1, 5.2, 5.3, 5.4 and 5.5). As shown in Table. 5.1 and 5.3, calculation of genetic correlation for traits with small sample sizes failed in those two bins, because the h^2 (heritability) estimates are negative in those cases. As genetic correlation between two traits are calculated using genetic covariance and h^2, we also provided the tables of genetic covariance and h^2 in the Supplement (Table. 5.6, 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12).



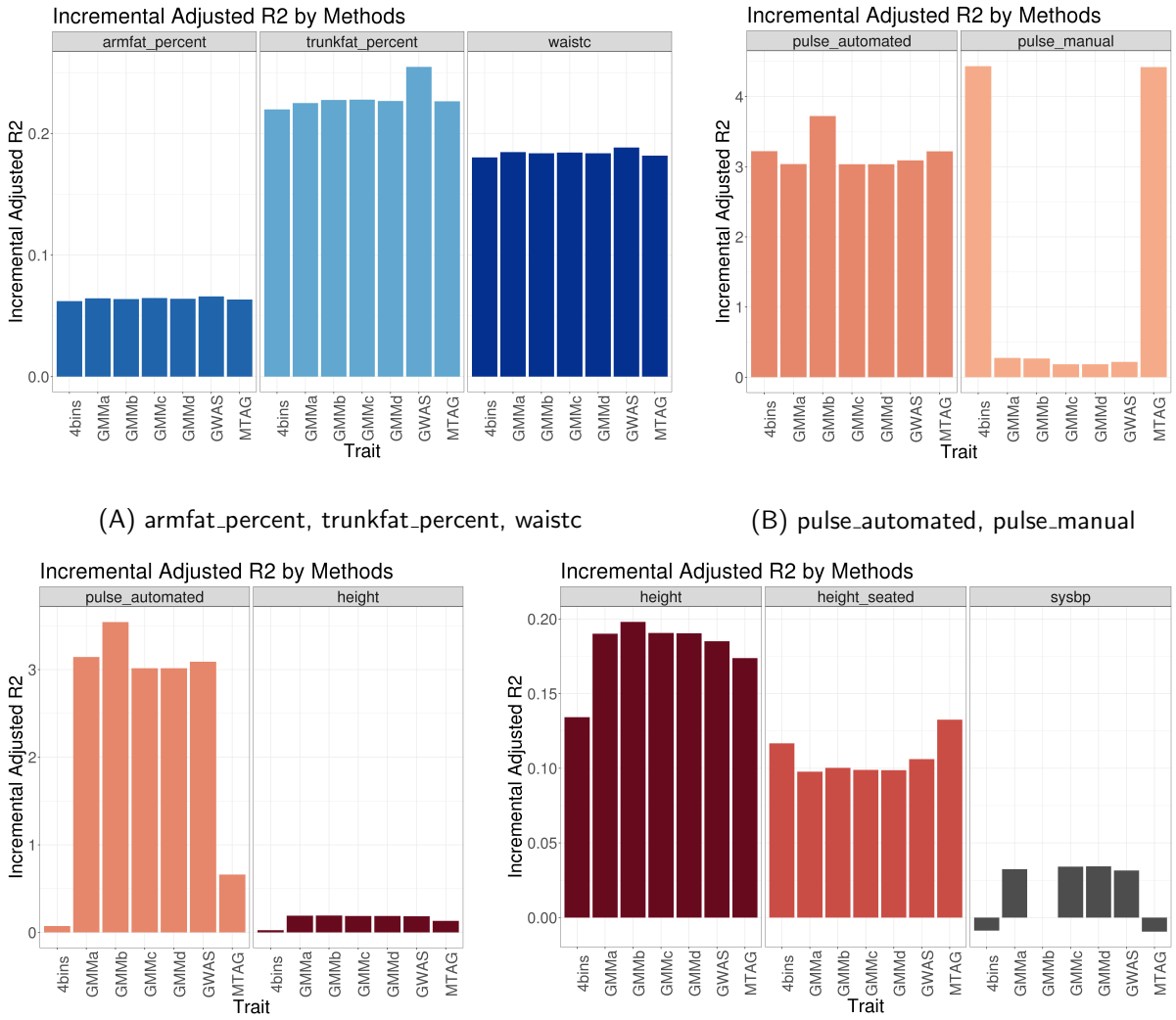
(A) armfat_percent, trunkfat_percent, waistc

(B) pulse_automated, pulse_manual

(C) pulse_automated, height

(D) height, height_seated, sysbp

Figure 3.1: Predictive power of mixture models, GWAS, and MTAG on four sets of anthropometric traits. Polygenic risk scores (PRS) were computed using one of five mixture models, GWAS, or MTAG effect size estimates. Incremental r^2 is proportion increase in r^2 between the PRS and observed phenotypes, compared to prediction using linear model with covariates only. Note that for sysbp, the PRS from GMM_B estimates did not run successfully during LDpred prediction step, because the mean χ^2 statistic is too small.



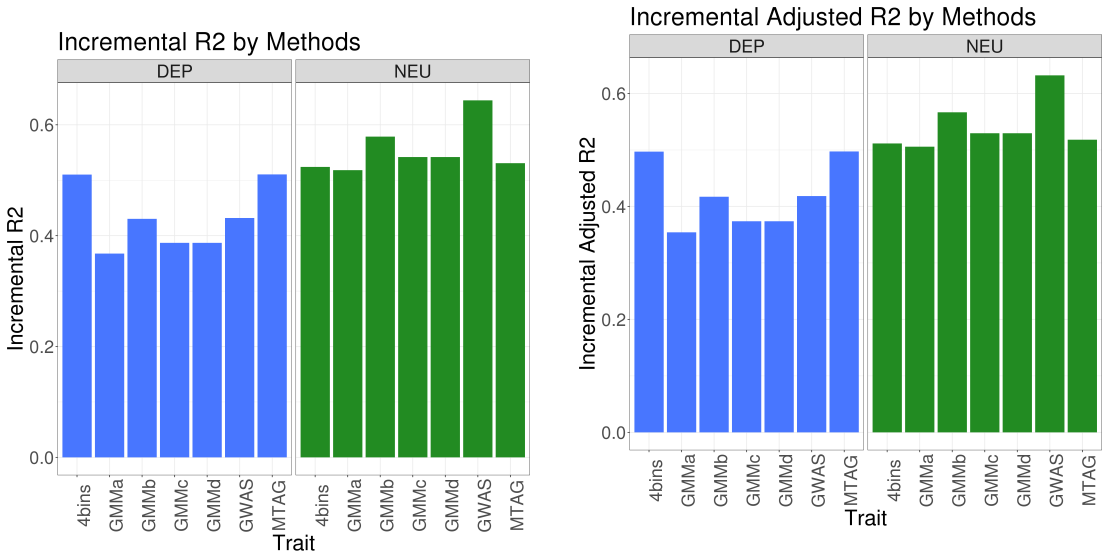
(A) armfat_percent, trunkfat_percent, waistc

(B) pulse_automated, pulse_manual

(C) pulse_automated, height

(D) height, height_seated, sysbp

Figure 3.2: Predictive power of mixture models, GWAS, and MTAG on four sets of anthropometric traits. Polygenic risk scores (PRS) were computed using one of five mixture models, GWAS, or MTAG effect size estimates. Adjusted incremental r^2 is incremental r^2 corrected for additional number of predictors compared to covariate-only linear model. Note that for sysbp, the PRS from GMM_B estimates did not run successfully during LDpred prediction step, because the mean χ^2 statistic is too small.



(A) Incremental R2 for depression and neuroticism

(B) Incremental adjusted R2 for depression and neuroticism

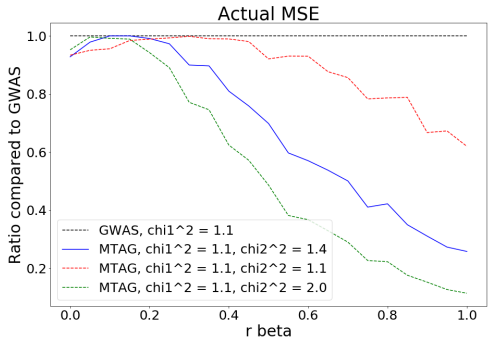
Figure 3.3: Predictive power of mixture models, GWAS, and MTAG on a set of psychiatric traits, depression and neuroticism. Polygenic risk scores (PRS) were computed using one of five mixture models, GWAS, or MTAG effect size estimates. Adjusted incremental r^2 is incremental r^2 corrected for additional number of predictors compared to covariate-only linear model.

3.2 MSE comparison in simulation

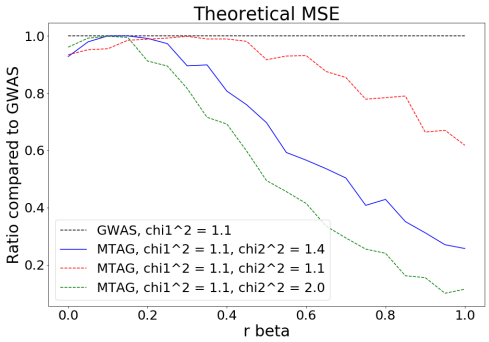
We generated the effect sizes, effect size distribution and estimation errors using the same parameter settings as Turley *et. al* and computed the theoretical mean squared error (MSE) using the same formula as Turley *et. al* for GWAS and MTAG [TW18] (Fig: 3.4A and 3.4B). We also computed the actual MSE for the GWAS, MTAG and GMM_A method for the same simulation data (Fig: 3.4C).

The MTAG simulation results replicate those shown in Turley *et. al* [TW18], showing a trend of lower MSE when there is higher genetic correlation among traits and when χ^2 is higher (indicating higher heritability and larger estimation sample size). The MSE's from

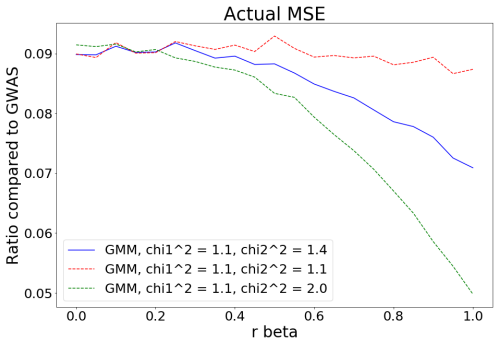
GMM method are much lower than the MSE's from GWAS method or MTAG method. This indicates that if our assumptions of effect size distribution are correct, the multi-trait methods should perform in the way that we expected, and that GMM method predicts with much higher accuracy than the other methods in simulated data.



(A) Ratio of actual MSE of MTAG estimates to MSE of GWAS estimates



(B) Ratio of theoretical MSE of MTAG estimates to MSE of GWAS estimates



(C) Ratio of actual MSE of GMM estimates to MSE of GWAS estimates

Figure 3.4: MSE computed from simulation results

Table 3.1: Genetic correlation (standard error) between UKBB anthropometric traits. LDSC calculated from the filtered SNPs

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**1.0000**	-0.152(0.025)	-0.089(0.031)	0.107(0.030)	0.039(0.070)	0.025(0.083)	0.947(0.004)	0.888(0.008)
height	-0.152(0.025)	**0.9998**	0.918(0.007)	-0.112(0.024)	-0.127(0.064)	-0.156(0.071)	0.055(0.026)	0.122(0.024)
height_s	-0.089(0.031)	0.918(0.007)	**0.9999**	-0.069(0.026)	-0.078(0.068)	-0.121(0.079)	0.090(0.032)	0.162(0.028)
pulse_a	0.107(0.030)	-0.112(0.024)	-0.069(0.026)	**1.0000**	1.028(0.138)	-0.028(0.103)	0.102(0.030)	0.071(0.031)
pulse_m	0.039(0.070)	-0.127(0.064)	-0.078(0.068)	1.028(0.138)	**1.0000**	-0.027(0.241)	-0.006(0.074)	-0.014(0.077)
sysbp	0.025(0.083)	0.156(0.071)	-0.121(0.079)	-0.028(0.103)	-0.027(0.241)	**1.0000**	-0.011(0.083)	-0.010(0.085)
trunkfat%	0.947(0.004)	0.055(0.026)	0.090(0.032)	0.102(0.030)	-0.006(0.074)	-0.011(0.083)	**1.0000**	0.863(0.008)
waistc	0.888(0.008)	0.122(0.024)	0.162(0.028)	0.071(0.031)	-0.014(0.077)	-0.010(0.085)	0.863(0.008)	**1.0000**

Table 3.2: List of traits used in the analysis and description

Trait abbreviation	Shorter abbreviation	Description
armfat_percent	armfat%	arm fat percentage, left arm
height	height	standing height
height_seated	height_s	seated height
pulse_automated	pulse_a	pulse rate, automated reading
pulse_manual	pulse_m	pulse rate, during blood-pressure measurement
sysbp	sysbp	systolic blood pressure, manual reading
trunkfat_percent	trunkfat%	trunk fat percentage
waistc	waistc	waist circumference
DEP	DEP	depression
NEU	NEU	neuroticism

Table 3.3: Number of final individual counts (sizes of GWAS cohort) for each anthropometric trait in the 200k randomly sampled individuals from UK Biobank

Trait	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
Num ind	196573	199573	199576	188947	18805	18805	196521	199671

Table 3.4: Number of final individual counts (sizes of GWAS cohort) for each psychiatric trait in the 200k randomly sampled individuals from UK Biobank

Trait	DEP	NEU
Num ind	180390	162694

Table 3.5: SNP bins used for the stratified method, partitioned by LD score and MAF (upper bounds exclusive) LDSC calculated from the filtered SNPs

Bin	LD score quartile range	MAF range	Number of SNPs
1	0 - 50	0 - 0.1	66071
2	50 - 100	0 - 0.1	38936
3	0 - 50	0.1 - 0.5	46544
4	50 - 100	0.1 - 0.5	73717

CHAPTER 4

Discussion

Through applying GWAS and various multi-trait approaches to UK Biobank anthropometric and psychiatric traits and simulated datasets, we aim to assess whether the SNP effect sizes estimated by any of the approaches would improve the predictive power of polygenic risk score.

From the UK Biobank results, we observe that the 4 bins model performs with a high similarity to MTAG than to GWAS or other multi-trait methods. We also observe that the GMM methods' performance have an overall similar pattern to GWAS method and other GMM methods, with GMM_B sometimes having a better performance than other GMM methods. We realize that in traits that are highly genetically correlated and when one of the traits has a small sample size, such as in the automated pulse and manual pulse set (Fig. 3.2B), both the 4 bins method and MTAG method boost the predictive power more strongly than other GMM methods for the trait. However, when the two traits have low or negative genetic correlation, such as in automated pulse and height set, and in standing height, seated height and systolic blood pressure set (Fig. 3.2C and 3.2D), 4 bins and MTAG methods have the lowest performance compared to other methods. In the cases where the traits have high genetic correlation and the sample sizes of the two traits are comparably large, we cannot determine which one of the methods has the highest performance, as GWAS method outperforms other methods in some cases (Fig. 3.2A), while the 4 bins and MTAG methods outperform in other cases (Fig. 3.3B). Since we tested on various sets of traits that represent different scenarios, including the set of traits that was analysed in Turley *et. al*

[TW18], we conclude that multi-trait method does not perform consistently better in all of the scenarios than GWAS.

We suspect the reason that 4 bins and MTAG methods are similar in performance could be because SNPs in one of the bins provide the most information and have relatively larger effect sizes, making the effect size estimates of that specific bin representative of the genome.

In order to further validate our conclusion, performing similar experiments using different sub-sample sizes and observing potential trends in each methods' performance with respect to sample size will provide useful insights. In addition, for the 4 bins method, performing a controlled experiment that involves random stratification of 4 bins, and comparing the PRS generated from SNPs from any of the 4 bins with the overall PRS could also be helpful.

CHAPTER 5

Supplement

5.1 LDSC genetic correlation in 4 bins model

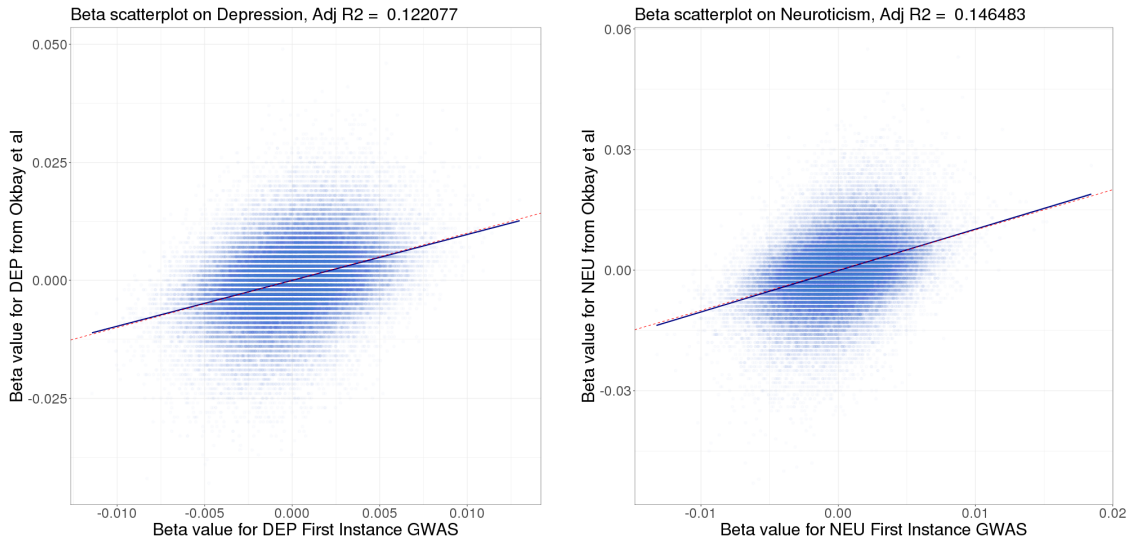
The genetic correlation matrices calculated in the 4 bins model are listed in (Table. 5.1, 5.2, 5.3, 5.4 and 5.5)

5.2 LDSC genetic covariance and heritability estimation for all SNPs and in 4 bins model

The genetic covariance and heritability tables calculated for all SNPs and for SNPs in the 4 bins model are listed in (Table. 5.6, 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12)

5.3 GWAS effect size estimates correlations for UKB psychiatric traits

The GWAS effect size estimates for depression and neuroticism are calculated from UKB psychiatric traits, and they are referred to as DEP first instance GWAS and NEU first instance GWAS. We compare those estimates with the GWAS effect size estimates for same traits using the large cohort analysed by Turley *et. al* [TW18] [LOB16]. The results are shown in Fig. 5.1.



(A) GWAS effect size estimates correlation between UKB depression dataset and depression dataset from previous group

(B) GWAS effect size estimates correlation between UKB neuroticism dataset and neuroticism dataset from previous group

Figure 5.1: GWAS effect size estimates correlations for UKB psychiatric traits. Red dashed line corresponds to $y = x$ line and blue line is the regression trend line for the dataset

5.4 Proportion of SNP effect size assignments in GMM models

The proportion of SNP effect size assignments for each trait in each GMM model after EM steps are shown in Table. 5.13. π_0 is the proportion of SNP assignments to distribution component with variance-covariance matrix $\mathbf{\Omega}_0$, and π_1 is the proportion of SNP effect size assignments to distribution component with variance-covariance matrix $\mathbf{\Omega}_1$.

Table 5.1: Genetic correlation (standard error) between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0-0.1 and LD score in the 0-50th percentile. LDSC calculated from the filtered SNPs

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**1.000**	-0.164(0.072)	-0.084(0.107)	0.199(0.110)	0.140(0.318)	NA	0.955(0.011)	0.879(0.027)
height	-0.164(0.072)	**1.000**	0.935(0.046)	-0.200(0.087)	-0.025(0.234)	NA	0.017(0.074)	0.112(0.072)
height_s	-0.084(0.107)	0.935(0.046)	**1.000**	-0.115(0.136)	0.148(0.401)	NA	0.058(0.108)	0.198(0.104)
pulse_a	0.199(0.110)	-0.200(0.087)	-0.115(0.136)	**1.000**	0.812(0.728)	NA	0.171(0.110)	0.077(0.110)
pulse_m	0.140(0.318)	-0.025(0.234)	0.149(0.401)	0.812(0.728)	**1.000**	NA	-0.018(0.297)	-0.198(0.332)
sysbp	NA	NA	NA	NA	NA	NA	NA	NA
trunkfat%	0.955(0.011)	0.017(0.074)	0.058(0.108)	0.171(0.110)	-0.018(0.297)	NA	**1.000**	0.846(0.031)
waistc	0.879(0.027)	0.112(0.072)	0.198(0.104)	0.077(0.110)	-0.198(0.332)	NA	0.846(0.031)	**1.000**

Table 5.2: Genetic correlation (standard error) between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0-0.1 and LD score in the 50-100th percentile. LDSC calculated from the filtered SNPs

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**1.000**	-0.128(0.077)	-0.125(0.083)	0.041(0.010)	0.157(0.158)	-0.063(0.216)	0.932(0.019)	0.872(0.026)
height	-0.128(0.077)	**1.000**	0.929(0.019)	-0.039(0.084)	-0.003(0.110)	0.005(0.174)	0.147(0.099)	0.218(0.095)
height_s	-0.125(0.083)	0.929(0.019)	**1.000**	0.018(0.090)	-0.045(0.127)	-0.013(0.205)	0.108(0.090)	0.167(0.091)
pulse_a	0.041(0.010)	-0.039(0.084)	0.018(0.090)	**1.000**	0.614(0.169)	-0.479(0.328)	0.041(0.102)	-0.009(0.109)
pulse_m	0.157(0.158)	-0.003(0.110)	-0.045(0.127)	0.614(0.169)	**1.000**	-0.264(0.381)	0.198(0.153)	0.027(0.160)
sysbp	-0.063(0.216)	0.005(0.174)	-0.013(0.205)	-0.479(0.328)	-0.264(0.381)	**1.000**	0.028(0.230)	-0.114(0.229)
trunkfat%	0.932(0.019)	0.147(0.099)	0.108(0.090)	0.041(0.102)	0.198(0.153)	0.028(0.230)	**1.000**	0.881(0.026)
waistc	0.872(0.026)	0.218(0.095)	0.167(0.091)	-0.009(0.109)	0.027(0.160)	-0.114(0.228)	0.881(0.026)	**1.000**

Table 5.3: Genetic correlation (standard error) between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0.1-0.5 and LD score in the 0-50th percentile. LDSC calculated from the filtered SNPs

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**1.000**	-0.152(0.050)	-0.166(0.067)	0.106(0.101)	NA	0.865(2.257)	0.969(0.010)	0.908(0.018)
height	-0.152(0.050)	**1.000**	0.950(0.030)	-0.114(0.077)	NA	-0.406(1.077)	0.054(0.056)	0.090(0.058)
height_s	-0.166(0.067)	0.950(0.030)	**1.000**	-0.180(0.122)	NA	-0.183(0.659)	0.024(0.072)	0.077(0.071)
pulse_a	0.106(0.101)	-0.114(0.077)	-0.180(0.122)	**1.000**	NA	0.813(2.163)	0.063(0.106)	0.132(0.010)
pulse_m	NA	NA	NA	NA	NA	NA	NA	NA
sysbp	0.865(2.257)	-0.406(1.077)	-0.183(0.659)	0.813(2.163)	NA	**1.000**	0.920(2.385)	1.003(2.591)
trunkfat%	0.969(0.010)	0.054(0.056)	0.024(0.072)	0.063(0.106)	NA	0.920(2.385)	**1.000**	0.881(0.025)
waistc	0.908(0.018)	0.090(0.058)	0.077(0.071)	0.132(0.010)	NA	1.003(2.591)	0.881(0.025)	**1.000**

Table 5.4: Genetic correlation (standard error) between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0.1-0.5 and LD score in the 50-100th percentile. LDSC calculated from the filtered SNPs

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**1.000**	-0.128(0.042)	-0.074(0.049)	0.025(0.054)	0.143(0.122)	0.056(0.168)	0.941(0.008)	0.890(0.016)
height	-0.128(0.042)	**1.000**	0.935(0.01)	-0.105(0.048)	-0.172(0.119)	-0.321(0.168)	0.081(0.046)	0.172(0.039)
height_s	-0.074(0.049)	0.935(0.01)	**1.000**	-0.105(0.051)	-0.135(0.127)	-0.298(0.182)	0.107(0.052)	0.198(0.044)
pulse_a	0.025(0.054)	-0.105(0.048)	-0.105(0.051)	**1.000**	1.172(0.245)	-0.089(0.213)	0.042(0.055)	-0.022(0.053)
pulse_m	0.143(0.122)	-0.172(0.119)	-0.135(0.127)	1.172(0.245)	**1.000**	-0.390(0.526)	0.133(0.120)	-0.012(0.125)
sysbp	0.056(0.168)	-0.321(0.168)	-0.298(0.182)	-0.089(0.213)	-0.39(0.526)	**1.000**	0.040(0.164)	-0.045(0.173)
trunkfat%	0.941(0.008)	0.081(0.046)	0.107(0.052)	0.042(0.055)	0.133(0.120)	0.040(0.164)	**1.000**	0.86(0.016)
waistc	0.890(0.016)	0.172(0.039)	0.198(0.044)	-0.022(0.053)	-0.012(0.125)	-0.045(0.173)	0.86(0.0163)	**1.000**

Table 5.5: Genetic correlation (standard error) table between UKBB psychiatric traits, stratified into 4 bins by LD score and MAF. LDSC calculated from the filtered SNPs.

(a) SNPs with MAF 0-0.1 and LD score in the 0-50th percentile.

	DEP	NEU
DEP	**1.000**	0.987(0.121)
NEU	0.987(0.121)	**1.000**

(b) SNPs with MAF 0-0.1 and LD score in the 50-100th percentile.

	DEP	NEU
DEP	**1.000**	0.896(0.074)
NEU	0.896(0.074)	**1.000**

(c) SNPs with MAF 0.1-0.5 and LD score in the 0-50th percentile.

	DEP	NEU
DEP	**1.000**	0.9(0.145)
NEU	0.9(0.145)	**1.000**

(d) SNPs with MAF 0.1-0.5 and LD score in the 50-100th percentile.

	DEP	NEU
DEP	**1.000**	0.778(0.052)
NEU	0.778(0.052)	**1.000**

Table 5.6: Genetic covariance and heritability (standard error) table between UKBB anthropometric traits (no bins). LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

	armfat_percent	height	height_seated	pulse_automated	pulse_manual	sysbp	trunkfat_percent	waistc
armfat_percent	**0.235(0.012)**	-0.054(0.009)	-0.021(0.007)	0.02(0.006)	0.006(0.012)	0.004(0.013)	0.213(0.01)	0.2(0.011)
height	-0.054(0.009)	**0.527(0.032)**	0.327(0.022)	-0.031(0.007)	-0.031(0.015)	-0.036(0.016)	0.019(0.009)	0.041(0.009)
height_seated	-0.021(0.007)	0.327(0.022)	**0.241(0.018)**	-0.013(0.005)	-0.013(0.011)	-0.019(0.012)	0.02(0.007)	0.037(0.007)
pulse_automated	0.02(0.006)	-0.031(0.007)	-0.013(0.005)	**0.147(0.014)**	0.133(0.017)	-0.003(0.013)	0.018(0.005)	0.013(0.006)
pulse_manual	0.006(0.012)	-0.031(0.015)	-0.013(0.011)	0.133(0.017)	**0.114(0.036)**	-0.003(0.026)	-0.001(0.012)	-0.002(0.012)
sysbp	0.004(0.013)	-0.036(0.016)	-0.019(0.012)	-0.003(0.013)	-0.003(0.026)	**0.103(0.038)**	-0.002(0.012)	-0.002(0.013)
trunkfat_percent	0.213(0.01)	0.019(0.009)	0.02(0.007)	0.018(0.005)	-0.001(0.012)	-0.002(0.012)	**0.216(0.01)**	0.186(0.01)
waistc	0.2(0.011)	0.041(0.009)	0.037(0.007)	0.013(0.006)	-0.002(0.012)	-0.002(0.013)	0.186(0.01)	**0.216(0.012)**

Table 5.7: Genetic covariance and heritability (standard error) table between UKBB psychiatric traits (no bins). LDSC calculated from the filtered SNPs

	DEP	NEU
DEP	**0.075(0.006)**	0.07(0.005)
NEU	0.07(0.005)	**0.1(0.007)**

Table 5.8: Genetic covariance and heritability (standard error) table between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0-0.1 and LD score in the 0-50th percentile. LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**0.159(0.019)**	-0.037(0.017)	-0.011(0.014)	0.025(0.014)	0.018(0.038)	-0.001(0.039)	0.152(0.017)	0.139(0.018)
height	-0.037(0.017)	**0.314(0.04)**	0.173(0.025)	-0.035(0.015)	-0.004(0.042)	0.015(0.039)	0.004(0.017)	0.025(0.016)
height_s	-0.011(0.014)	0.173(0.025)	**0.108(0.022)**	-0.012(0.014)	0.016(0.04)	0.024(0.037)	0.008(0.014)	0.026(0.014)
pulse_a	0.025(0.014)	-0.035(0.015)	-0.012(0.014)	**0.098(0.021)**	0.082(0.046)	-0.025(0.041)	0.021(0.014)	0.01(0.014)
pulse_m	0.018(0.038)	-0.004(0.042)	0.016(0.04)	0.082(0.046)	**0.104(0.178)**	-0.015(0.112)	-0.002(0.038)	0.025(0.037)
sysbp	-0.001(0.039)	0.015(0.039)	0.024(0.037)	-0.025(0.041)	-0.015(0.112)	**-0.014(0.16)**	0(0.037)	0.017(0.037)
trunkfat%	0.152(0.017)	0.004(0.017)	0.008(0.014)	0.021(0.014)	-0.002(0.038)	0(0.037)	**0.16(0.018)**	0.134(0.017)
waistc	0.139(0.018)	0.025(0.016)	0.026(0.014)	0.01(0.014)	0.025(0.037)	0.017(0.037)	0.134(0.017)	**0.158(0.02)**

Table 5.9: Genetic covariance and heritability (standard error) table between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0-0.1 and LD score in the 50-100th percentile. LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**0.141(0.026)**	-0.034(0.02)	-0.022(0.014)	0.005(0.012)	0.031(0.03)	-0.008(0.025)	0.132(0.023)	0.115(0.022)
height	-0.034(0.02)	**0.495(0.09)**	0.307(0.048)	-0.009(0.019)	-0.001(0.04)	0.001(0.038)	0.039(0.028)	0.054(0.027)
height_s	-0.022(0.014)	0.307(0.048)	**0.221(0.033)**	0.003(0.014)	-0.011(0.031)	-0.002(0.03)	0.019(0.016)	0.028(0.017)
pulse_a	0.005(0.012)	-0.009(0.019)	0.003(0.014)	**0.113(0.028)**	0.109(0.032)	-0.052(0.034)	0.005(0.013)	-0.001(0.012)
pulse_m	0.031(0.03)	-0.001(0.04)	-0.011(0.031)	0.109(0.032)	**0.277(0.09)**	-0.044(0.059)	0.039(0.028)	0.005(0.029)
sysbp	-0.008(0.025)	0.001(0.038)	-0.002(0.03)	-0.052(0.034)	-0.044(0.059)	**0.102(0.073)**	0.003(0.027)	-0.013(0.025)
trunkfat%	0.132(0.023)	0.039(0.028)	0.019(0.016)	0.005(0.013)	0.039(0.028)	0.003(0.027)	**0.141(0.024)**	0.116(0.022)
waistc	0.115(0.022)	0.054(0.027)	0.028(0.017)	-0.001(0.012)	0.005(0.029)	-0.013(0.025)	0.116(0.022)	**0.123(0.024)**

Table 5.10: Genetic covariance and heritability (standard error) table between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0.1-0.5 and LD score in the 0-50th percentile. LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**0.275(0.026)**	-0.059(0.02)	-0.042(0.017)	0.019(0.018)	0.02(0.05)	0.097(0.05)	0.259(0.024)	0.24(0.023)
height	-0.059(0.02)	**0.558(0.067)**	0.342(0.041)	-0.029(0.02)	-0.046(0.057)	-0.065(0.059)	0.02(0.022)	0.034(0.022)
height_s	-0.042(0.017)	0.342(0.041)	**0.232(0.03)**	-0.029(0.018)	-0.035(0.049)	-0.019(0.051)	0.006(0.018)	0.019(0.018)
pulse_a	0.019(0.018)	-0.029(0.02)	-0.029(0.018)	**0.114(0.038)**	0.103(0.055)	0.058(0.052)	0.011(0.018)	0.022(0.017)
pulse_m	0.02(0.05)	-0.046(0.057)	-0.035(0.049)	0.103(0.055)	**-0.009(0.21)**	-0.036(0.145)	0.013(0.048)	0.068(0.047)
sysbp	0.097(0.05)	-0.065(0.059)	-0.019(0.051)	0.058(0.052)	-0.036(0.145)	**0.046(0.183)**	0.1(0.051)	0.108(0.048)
trunkfat%	0.259(0.024)	0.02(0.022)	0.006(0.018)	0.011(0.018)	0.013(0.048)	0.1(0.051)	**0.26(0.026)**	0.226(0.022)
waistc	0.24(0.023)	0.034(0.022)	0.019(0.018)	0.022(0.017)	0.068(0.047)	0.108(0.048)	0.226(0.022)	**0.253(0.026)**

Table 5.11: Genetic covariance and heritability (standard error) table between UKBB anthropometric traits, stratified into 4 bins by LD score and MAF. For SNPs with MAF 0.1-0.5 and LD score in the 50-100th percentile. LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

	armfat%	height	height_s	pulse_a	pulse_m	sysbp	trunkfat%	waistc
armfat%	**0.294(0.033)**	-0.06(0.02)	-0.023(0.015)	0.005(0.011)	0.027(0.023)	0.008(0.024)	0.261(0.028)	0.256(0.034)
height	-0.06(0.02)	**0.758(0.072)**	0.465(0.05)	-0.036(0.016)	-0.052(0.034)	-0.075(0.033)	0.036(0.021)	0.079(0.02)
height_s	-0.023(0.015)	0.465(0.05)	**0.327(0.039)**	-0.024(0.012)	-0.027(0.025)	-0.046(0.024)	0.032(0.015)	0.06(0.015)
pulse_a	0.005(0.011)	-0.036(0.016)	-0.024(0.012)	**0.155(0.028)**	0.16(0.033)	-0.009(0.022)	0.008(0.011)	-0.005(0.011)
pulse_m	0.027(0.023)	-0.052(0.034)	-0.027(0.025)	0.16(0.033)	**0.12(0.064)**	-0.036(0.04)	0.024(0.022)	-0.002(0.022)
sysbp	0.008(0.024)	-0.075(0.033)	-0.046(0.024)	-0.009(0.022)	-0.036(0.04)	**0.071(0.055)**	0.006(0.022)	-0.006(0.024)
trunkfat%	0.261(0.028)	0.036(0.021)	0.032(0.015)	0.008(0.011)	0.024(0.022)	0.006(0.022)	**0.262(0.026)**	0.233(0.028)
waistc	0.256(0.034)	0.079(0.02)	0.06(0.015)	-0.005(0.011)	-0.002(0.022)	-0.006(0.024)	0.233(0.028)	**0.281(0.037)**

Table 5.12: Genetic covariance and heritability (standard error) table between UKBB psychiatric traits, stratified into 4 bins by LD score and MAF. LDSC calculated from the filtered SNPs. Diagonal entries are heritability estimates for each trait, and off-diagonal entries are genetic covariance estimates for each pair of traits.

(a) SNPs with MAF 0-0.1 and LD score in the 0-50th percentile.

	DEP	NEU
DEP	**0.065(0.018)**	0.064(0.016)
NEU	0.064(0.016)	**0.066(0.021)**

(b) SNPs with MAF 0-0.1 and LD score in the 50-100th percentile.

	DEP	NEU
DEP	**0.06(0.013)**	0.056(0.011)
NEU	0.056(0.011)	**0.066(0.012)**

(c) SNPs with MAF 0.1-0.5 and LD score in the 0-50th percentile.

	DEP	NEU
DEP	**0.052(0.021)**	0.072(0.019)
NEU	0.072(0.019)	**0.125(0.025)**

(d) SNPs with MAF 0.1-0.5 and LD score in the 50-100th percentile.

	DEP	NEU
DEP	**0.072(0.009)**	0.061(0.009)
NEU	0.061(0.009)	**0.086(0.013)**

(a) Proportion of assignments in GMM models for anthropometric trait set (1)

Model	GMM_A		GMM_B		GMM_C		GMM_D	
Assignments	π_0	π_1	π_0	π_1	π_0	π_1	π_0	π_1
armfat%	0.006	0.994	0.022	0.978	0.006	0.994	0.008	0.992
trunkfat%	0.049	0.951	0.004	0.996	0.008	0.992	0.008	0.992
waistc	0.061	0.939	0.021	0.979	0.007	0.993	0.008	0.992

(b) Proportion of assignments in GMM models for anthropometric trait set (2)

Model	GMM_A		GMM_B		GMM_C		GMM_D	
Assignments	π_0	π_1	π_0	π_1	π_0	π_1	π_0	π_1
pulse_a	0.236	0.764	0.953	0.047	0.044	0.956	0.044	0.956
pulse_m	0.953	0.047	0.236	0.764	0.044	0.956	0.044	0.956

(c) Proportion of assignments in GMM models for anthropometric trait set (3)

Model	GMM_A		GMM_B		GMM_C		GMM_D	
Assignments	π_0	π_1	π_0	π_1	π_0	π_1	π_0	π_1
pulse_a	0.830	0.170	0.880	0.120	0.614	0.386	0.614	0.386
height	0.880	0.120	0.830	0.170	0.614	0.386	0.614	0.386

(d) Proportion of assignments in GMM models for anthropometric trait set (4)

Model	GMM_A		GMM_B		GMM_C		GMM_D	
Assignments	π_0	π_1	π_0	π_1	π_0	π_1	π_0	π_1
height	0.830	0.170	0.880	0.120	0.614	0.386	0.614	0.386
height_s	0.880	0.120	0.830	0.170	0.614	0.386	0.614	0.386
sysbp	0.929	0.071	0.988	0.012	0.214	0.786	0.001	0.999

(e) Proportion of assignments in GMM models for psychiatric trait set

Model	GMM_A		GMM_B		GMM_C		GMM_D	
Assignments	π_0	π_1	π_0	π_1	π_0	π_1	π_0	π_1
DEP	0.292	0.708	0.424	0.576	0.047	0.953	0.047	0.953
NEU	0.424	0.576	0.292	0.708	0.047	0.953	0.047	0.953

Table 5.13: Proportions of component assignments in four GMM models for all of the trait sets

REFERENCES

[BF15] Brendan Bulik-Sullivan, , Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price, Benjamin M Neale, and and. "An atlas of genetic correlations across human diseases and traits." *Nature Genetics*, **47**(11):1236–1241, sep 2015.

[BL15] Brendan K Bulik-Sullivan, , Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nature Genetics*, **47**(3):291–295, feb 2015.

[GRA17] Jakob Grove, Stephan Ripke, Thomas Damm Als, Manuel Mattheisen, Raymond Walters, Hyejung Won, Jonatan Pallesen, Esben Agerbo, Ole A Andreassen, Richard Anney, Rich Belliveau, Francesco Bettella, Joseph D Buxbaum, Jonas Bybjerg-Grauholm, Marie Bækved-Hansen, Felecia Cerrato, Kimberly Chambert, Jane H Christensen, Claire Churchhouse, Karin Dellenvall, Ditte Demontis, Silvia De Rubeis, Bernie Devlin, Srdjan Djurovic, Ashley Dumont, Jacqueline Goldstein, Christine S Hansen, Mads E Hauberg, Mads V Hollegaard, Sigrun Hope, Daniel P Howrigan, Hailiang Huang, Christina Hultman, Lambertus Klei, Julian Maller, Joanna Martin, Alicia R Martin, Jennifer Moran, Mette Nyegaard, Terje Nærland, Duncan S Palmer, Aarno Palotie, Carsten B Pedersen, Marianne G Pedersen, Timothy Poterba, Jesper B Poulsen, Beate St Pourcain, Per Qvist, Karola Rehnström, Avi Reichenberg, Jennifer Reichert, Elise Robinson, Kathryn Roeder, Panos Roussos, Evald Saemundsen, Sven Sandin, F Kyle Satterstrom, George D Smith, Hreinn Stefansson, Kari Stefansson, Stacy Steinberg, Christine Stevens, Patrick F Sullivan, Patrick Turley, G Bragi Walters, Xinyi Xu, Daniel Geschwind, Merete Nordentoft, David M Hougaard, Thomas Werge, Ole Mors, Preben B Mortensen, Benjamin M Neale, Mark J Daly, Anders D Børglum, , , and and. "Common risk variants identified in autism spectrum disorder." nov 2017.

[HLL17] Yiming Hu, Qiongshi Lu, Wei Liu, Yuhua Zhang, Mo Li, and Hongyu Zhao. "Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction." *PLOS Genetics*, **13**(6):e1006836, jun 2017.

[LOB16] LifeLines Cohort Study, Aysu Okbay, Bart M L Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G Nivard, Mark Alan Fontana, S Fleur W Meddens, Richard Karlsson Linnér, Cornelius A Rietveld, Jaime Derringer, Jacob Gratten, James J Lee, Jimmy Z Liu, Ronald de Vlaming, Tarunveer S Ahluwalia, Jadwiga Buchwald, Alana Cavadino, Alexis C Frazier-Wood, Nicholas A Furlotte, Victoria Garfield, Marie Henrike Geisel, Juan R Gonzalez, Saskia Haitjema, Robert Karlsson, Sander W van der Laan, Karl-Heinz Ladwig, Jari Lahti, Sven J van der

Lee, Penelope A Lind, Tian Liu, Lindsay Matteson, Evelin Mihailov, Michael B Miller, Camelia C Minica, Ilja M Nolte, Dennis Mook-Kanamori, Peter J van der Most, Christopher Oldmeadow, Yong Qian, Olli Raitakari, Rajesh Rawal, Anu Realo, Rico Rueedi, Börge Schmidt, Albert V Smith, Evie Stergiakouli, Toshiko Tanaka, Kent Taylor, Gudmar Thorleifsson, Juho Wedenoja, Juergen Wellmann, Harm-Jan Westra, Sara M Willems, Wei Zhao, Najaf Amin, Andrew Bakshi, Sven Bergmann, Gyda Bjornsdottir, Patricia A Boyle, Samantha Cherney, Simon R Cox, Gail Davies, Oliver S P Davis, Jun Ding, Nese Direk, Peter Eibich, Rebecca T Emeny, Ghazaleh Fatemifar, Jessica D Faul, Luigi Ferrucci, Andreas J Forstner, Christian Gieger, Richa Gupta, Tamara B Harris, Juliette M Harris, Elizabeth G Holliday, Jouke-Jan Hottenga, Philip L De Jager, Marika A Kaakinen, Eero Kajantie, Ville Karhunen, Ivana Kolcic, Meena Kumari, Lenore J Launer, Lude Franke, Ruifang Li-Gao, David C Liewald, Marisa Koini, Anu Loukola, Pedro Marques-Vidal, Grant W Montgomery, Miriam A Mosing, Lavinia Paternoster, Alison Pattie, Katja E Petrovic, Laura Pulkki-Råback, Lydia Quaye, Katri Räikkönen, Igor Rudan, Rodney J Scott, Jennifer A Smith, Angelina R Sutin, Maciej Trzaskowski, Anna E Vinkhuyzen, Lei Yu, Delilah Zabaneh, John R Attia, David A Bennett, Klaus Berger, Lars Bertram, Dorret I Boomsma, Harold Snieder, Shun-Chiao Chang, Francesco Cucca, Ian J Deary, Cornelia M van Duijn, Johan G Eriksson, Ute Bültmann, Eco J C de Geus, Patrick J F Groenen, Vilmundur Gudnason, Torben Hansen, Catharine A Hartman, Claire M A Haworth, Caroline Hayward, Andrew C Heath, David A Hinds, Elina Hyppönen, William G Iacono, Marjo-Riitta Järvelin, Karl-Heinz Jöckel, Jaakko Kaprio, Sharon L R Kardia, Liisa Keltikangas-Järvinen, Peter Kraft, Laura D Kubzansky, Terho Lehtimäki, Patrik K E Magnusson, Nicholas G Martin, Matt McGue, Andres Metspalu, Melinda Mills, Renée de Mutsert, Albertine J Oldehinkel, Gerard Pasterkamp, Nancy L Pedersen, Robert Plomin, Ozren Polasek, Christine Power, Stephen S Rich, Frits R Rosendaal, Hester M den Ruijter, David Schlessinger, Helena Schmidt, Rauli Svento, Reinhold Schmidt, Behrooz Z Alizadeh, Thorkild I A Sørensen, Tim D Spector, John M Starr, Kari Stefansson, Andrew Steptoe, Antonio Terracciano, Unnur Thorsteinsdottir, A Roy Thurik, Nicholas J Timpson, Henning Tiemeier, André G Uitterlinden, Peter Vollenweider, Gert G Wagner, David R Weir, Jian Yang, Dalton C Conley, George Davey Smith, Albert Hofman, Magnus Johannesson, David I Laibson, Sarah E Medland, Michelle N Meyer, Joseph K Pickrell, Tõnu Esko, Robert F Krueger, Jonathan P Beauchamp, Philipp D Koellinger, Daniel J Benjamin, Meike Bartels, and David Cesarini. "Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses." *Nature Genetics*, **48**(6):624–633, June 2016.

[LPE18] Mark W. Logue, Matthew S. Panizzon, Jeremy A. Elman, Nathan A. Gillespie, Sean N. Hatton, Daniel E. Gustavson, Ole A. Andreassen, Anders M. Dale, Carol E. Franz, Michael J. Lyons, Michael C. Neale, Chandra A. Reynolds, Xin

Tu, and William S. Kremen. "Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s." *Molecular Psychiatry*, feb 2018.

[LTY17] Max Lam, Joey W. Trampush, Jin Yu, Emma Knowles, Gail Davies, David C. Liewald, John M. Starr, Srdjan Djurovic, Ingrid Melle, Kjetil Sundet, Andrea Christoforou, Ivar Reinvang, Pamela DeRosse, Astri J. Lundervold, Vidar M. Steen, Thomas Espeseth, Katri Räikkönen, Elisabeth Widen, Aarno Palotie, Johan G. Eriksson, Ina Giegling, Bettina Konte, Panos Roussos, Stella Giakoumaki, Katherine E. Burdick, Antony Payton, William Ollier, Ornit Chiba-Falek, Deborah K. Attix, Anna C. Need, Elizabeth T. Cirulli, Aristotle N. Voineskos, Nikos C. Stefanis, Dimitrios Avramopoulos, Alex Hatzimanolis, Dan E. Arking, Nikolaos Smyrnis, Robert M. Bilder, Nelson A. Freimer, Tyrone D. Cannon, Edythe London, Russell A. Poldrack, Fred W. Sabb, Eliza Congdon, Emily Drabant Conley, Matthew A. Scult, Dwight Dickinson, Richard E. Straub, Gary Donohoe, Derek Morris, Aiden Corvin, Michael Gill, Ahmad R. Hariri, Daniel R. Weinberger, Neil Pendleton, Panos Bitsios, Dan Rujescu, Jari Lahti, Stephanie Le Hellard, Matthew C. Keller, Ole A. Andreassen, Ian J. Deary, David C. Glahn, Anil K. Malhotra, and Todd Lencz. "Large-Scale Cognitive GWAS Meta-Analysis Reveals Tissue-Specific Neural Expression and Potential Nootropic Drug Targets." *Cell Reports*, **21**(9):2597–2613, nov 2017.

[MZL18] Robert M. Maier, Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, Naomi R. Wray, Jian Yang, Peter M. Visscher, and Matthew R. Robinson. "Improving genetic prediction by leveraging genetic correlations among human diseases and traits." *Nature Communications*, **9**(1), mar 2018.

[PNT07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics*, **81**(3):559–575, September 2007.

[Pur20] Shaun Purcell. *PLINK (version 1.90)*, 2007 (accessed Jun 9, 2020). https://www.cog-genomics.org/plink/.

[QC17] Guanghao Qi and Nilanjan Chatterjee. "Heritability Informed Power Optimization (HIPO) Leads to Enhanced Detection of Genetic Associations Across Multiple Traits." *biorxiv*, nov 2017.

[SHM16] Yiwey Shieh, Donglei Hu, Lin Ma, Scott Huntsman, Charlotte C. Gard, Jessica W. T. Leung, Jeffrey A. Tice, Celine M. Vachon, Steven R. Cummings, Karla Kerlikowske, and Elad Ziv. "Breast cancer risk prediction using a clinical risk model

and polygenic risk score." *Breast Cancer Research and Treatment*, **159**(3):513–525, aug 2016.

[SMS17] Huwenbo Shi, Nicholas Mancuso, Sarah Spendlove, and Bogdan Pasaniuc. "Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits." *The American Journal of Human Genetics*, **101**(5):737–751, nov 2017.

[TW18] Patrick Turley, , Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A. Furlotte, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, David Laibson, David Cesarini, Benjamin M. Neale, and Daniel J. Benjamin and. "Multi-trait analysis of genome-wide association summary statistics using MTAG." *Nature Genetics*, **50**(2):229–237, jan 2018.

[VYF15] Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, Alkes L. Price, Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberly D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Lynn E. DeLisi, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Elliot S. Gershon, Ina Giegling, Paola Giusti-Rodrguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Jakob Grove, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann,

Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julia, Rene S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, Brian J. Kelly, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kahler, Claudine Laurent, Jimmy Lee Chee Keong, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lnnqvist, Milan Macek, Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Preben B. Mortensen, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Mller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietilinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Sderman, Srinivas Thirumalai, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Jing Qin Wu, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Rolf Adolfsson, Ole A. Andreassen, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tonu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jonsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levin-

son, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St. Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, Michael C. O'Donovan, Peter Kraft, David J. Hunter, Muriel Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Sonja Berndt, Carl Blomquist, Federico Canzian, Jenny Chang-Claude, Stephen J. Chanock, Laura Crisponi, Kamila Czene, Norbert Dahmen, Isabel dos Santos Silva, Douglas Easton, A. Heather Eliassen, Jonine Figueroa, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Lorna Gibson, Christopher A. Haiman, Per Hall, Aditi Hazra, Rebecca Hein, Brian E. Henderson, Albert Hofman, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Sara Lindström, Jianjun Liu, Eiliv Lund, Enes Makalic, Alfons Meindl, Hanne Meijers-Heijboer, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, María José Sánchez, Daniel F. Schmidt, Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth Travis, Clare Turnbull, Andre G. Uitterlinden, Rob B. van der Luijt, Quinten Waisfisz, Zhaoming Wang, Alice S. Whittemore, Rose Yang, and Wei Zheng. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics*, **97**(4):576–592, oct 2015.